jc685 U.S. PTO
10/24/00

Please type a plus sign (+) inside this box → [+]

PTO/SB/05 (4/98)
Approved for use through 09/30/2000. OMB 0651-0032
Patent and Trademark Office: U.S. DEPARTMENT OF COMMERCE

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

UTILITY PATENT APPLICATION TRANSMITTAL

(Only for new nonprovisional applications under 37 C.F.R. § 1.53(b))

Attorney Docket No.	219.39026X00
First Inventor or Application Identifier	Rajesh Shah
Title	See 1 in Addendum
Express Mail Label No.	

09/694492 PTO
10/24/00

APPLICATION ELEMENTS

See MPEP chapter 600 concerning utility patent application contents.

ADDRESS TO: Assistant Commissioner for Patents
Box Patent Application
Washington, DC 20231

1. [X] * Fee Transmittal Form *(e.g., PTO/SB/17)*
(Submit an original and a duplicate for fee processing)

2. [X] Specification [Total Pages 30]
(preferred arrangement set forth below)
- Descriptive title of the Invention
- Cross References to Related Applications
- Statement Regarding Fed sponsored R & D
- Reference to Microfiche Appendix
- Background of the Invention
- Brief Summary of the Invention
- Brief Description of the Drawings *(if filed)*
- Detailed Description
- Claim(s)
- Abstract of the Disclosure

3. [X] Drawing(s) *(35 U.S.C. 113)* [Total Sheets 7]

4. Oath or Declaration [Total Pages 3]
a. [X] Newly executed (original or copy)
b. [ ] Copy from a prior application (37 C.F.R. § 1.63(d))
(for continuation/divisional with Box 16 completed)
i. [ ] DELETION OF INVENTOR(S)
Signed statement attached deleting inventor(s) named in the prior application, see 37 C.F.R. §§ 1.63(d)(2) and 1.33(b).

** NOTE FOR ITEMS 1 & 13 IN ORDER TO BE ENTITLED TO PAY SMALL ENTITY FEES, A SMALL ENTITY STATEMENT IS REQUIRED (37 C.F.R. § 1.27), EXCEPT IF ONE FILED IN A PRIOR APPLICATION IS RELIED UPON (37 C.F.R. § 1.28).*

5. [ ] Microfiche Computer Program *(Appendix)*

6. Nucleotide and/or Amino Acid Sequence Submission
(if applicable, all necessary)
a. [ ] Computer Readable Copy
b. [ ] Paper Copy (identical to computer copy)
c. [ ] Statement verifying identity of above copies

ACCOMPANYING APPLICATION PARTS

7. [X] Assignment Papers (cover sheet & document(s))
8. [ ] 37 C.F.R.§3.73(b) Statement *(when there is an assignee)* [ ] Power of Attorney
9. [ ] English Translation Document *(if applicable)*
10. [ ] Information Disclosure Statement (IDS)/PTO-1449 [ ] Copies of IDS Citations
11. [ ] Preliminary Amendment
12. [X] Return Receipt Postcard (MPEP 503)
(Should be specifically itemized)
13. [ ] * Small Entity Statement(s) *(PTO/SB/09-12)* [ ] Statement filed in prior application Status still proper and desired
14. [ ] Certified Copy of Priority Document(s)
(if foreign priority is claimed)
15. [ ] Other: .....
.....
.....

16. If a CONTINUING APPLICATION, *check appropriate box, and supply the requisite information below and in a preliminary amendment:*
[ ] Continuation [ ] Divisional [ ] Continuation-in-part (CIP) of prior application No: _____/_____
Prior application information: Examiner _____ *Group / Art Unit:* _____
For CONTINUATION or DIVISIONAL APPS only: The entire disclosure of the prior application, from which an oath or declaration is supplied under Box 4b, is considered a part of the disclosure of the accompanying continuation or divisional application and is hereby incorporated by reference. The incorporation can only be relied upon when a portion has been inadvertently omitted from the submitted application parts.

17. CORRESPONDENCE ADDRESS

[X] *Customer Number or Bar Code Label* 020457 *(Insert Customer No. or Attach bar code label here)* or [ ] *Correspondence address below*

Name					
Address					
City		*State*		*Zip Code*	
Country		*Telephone*		*Fax*	

Name (Print/Type)	Christopher J. Hamaty	*Registration No. (Attorney/Agent)*	37,634
Signature		*Date*	10-24-00

Burden Hour Statement: This form is estimated to take 0.2 hours to complete. Time will vary depending upon the needs of the individual case. Any comments on the amount of time you are required to complete this form should be sent to the Chief Information Officer, Patent and Trademark Office, Washington, DC 20231. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Assistant Commissioner for Patents, Box Patent Application, Washington, DC 20231.

Attachment to PTO/SB/05 (4/98) Utility Patent Application Transmittal

1. SYSTEM AND METHOD FOR PROVIDING DETAILED PATH INFORATION TO CLIENTS.

PTO/SB/17 (12/99)
Approved for use through 09/30/2000. OMB 0651-0032
Patent and Trademark Office: U S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number

FEE TRANSMITTAL for FY 2000

Patent fees are subject to annual revision.
Small Entity payments must be supported by a small entity statement, otherwise large entity fees must be paid. See Forms PTO/SB/09-12 See 37 C.F.R. §§ 1.27 and 1 28.

TOTAL AMOUNT OF PAYMENT ($)786.00

Complete if Known	
Application Number	
Filing Date	October 24, 2000
First Named Inventor	Rajesh Shah
Examiner Name	
Group / Art Unit	
Attorney Docket No.	219.39026X00

METHOD OF PAYMENT (check one)

1. ☐ The Commissioner is hereby authorized to charge indicated fees and credit any overpayments to:

Deposit Account Number:

Deposit Account Name: 01-2135

☒ Charge Any Additional Fee Required Under 37 CFR §§ 1.16 and 1 17

2. ☒ **Payment Enclosed:**
☐ Check ☐ Money Order ☒ Other

FEE CALCULATION

1. BASIC FILING FEE

Large Entity Fee Code	Large Entity Fee ($)	Small Entity Fee Code	Small Entity Fee ($)	Fee Description	Fee Paid
101	690	201	345	Utility filing fee	710.00
106	310	206	155	Design filing fee	
107	480	207	240	Plant filing fee	
108	690	208	345	Reissue filing fee	
114	150	214	75	Provisional filing fee	

SUBTOTAL (1) ($) 710.00

2. EXTRA CLAIM FEES

			Extra Claims		Fee from below	=	Fee Paid
Total Claims	22	-20** =	2	X	18	=	36
Independent Claims	3	- 3** =	0	X	80	=	0
Multiple Dependent						=	0

***or number previously paid, if greater; For Reissues, see below*

Large Entity Fee Code	Large Entity Fee ($)	Small Entity Fee Code	Small Entity Fee ($)	Fee Description
103	18	203	9	Claims in excess of 20
102	78	202	39	Independent claims in excess of 3
104	260	204	130	Multiple dependent claim, if not paid
109	78	209	39	** Reissue independent claims over original patent
110	18	210	9	** Reissue claims in excess of 20 and over original patent

SUBTOTAL (2) ($) 36.00

FEE CALCULATION (continued)

3. ADDITIONAL FEES

Large Entity Fee Code	Large Entity Fee ($)	Small Entity Fee Code	Small Entity Fee ($)	Fee Description	Fee Paid
105	130	205	65	Surcharge - late filing fee or oath	0.00
127	50	227	25	Surcharge - late provisional filing fee or cover sheet.	0.00
139	130	139	130	Non-English specification	0.00
147	2,520	147	2,520	For filing a request for reexamination	0.00
112	920*	112	920*	Requesting publication of SIR prior to Examiner action	0.00
113	1,840*	113	1,840*	Requesting publication of SIR after Examiner action	0.00
115	110	215	55	Extension for reply within first month	0.00
116	380	216	190	Extension for reply within second month	0.00
117	870	217	435	Extension for reply within third month	0.00
118	1,360	218	680	Extension for reply within fourth month	0.00
128	1,850	228	925	Extension for reply within fifth month	0.00
119	300	219	150	Notice of Appeal	0.00
120	300	220	150	Filing a brief in support of an appeal	0.00
121	260	221	130	Request for oral hearing	0.00
138	1,510	138	1,510	Petition to institute a public use proceeding	0.00
140	110	240	55	Petition to revive - unavoidable	0.00
141	1,210	241	605	Petition to revive - unintentional	0.00
142	1,210	242	605	Utility issue fee (or reissue)	0.00
143	430	243	215	Design issue fee	0.00
144	580	244	290	Plant issue fee	0.00
122	130	122	130	Petitions to the Commissioner	0.00
123	50	123	50	Petitions related to provisional applications	0.00
126	240	126	240	Submission of Information Disclosure Stmt	0.00
581	40	581	40	Recording each patent assignment per property (times number of properties)	40.00
146	690	246	345	Filing a submission after final rejection (37 CFR § 1.129(a))	0.00
149	690	249	345	For each additional invention to be examined (37 CFR § 1 129(b))	0.00
				Other fee (specify)	0.00
				Other fee (specify)	0.00

* Reduced by Basic Filing Fee Paid

SUBTOTAL (3) ($) 40.00

SUBMITTED BY

Complete *(if applicable)*

Name (Print/Type)	Christopher J. Hamaty	Registration No. (Attorney/Agent)	37,634	Telephone	
Signature	Chris Hamaty			Date	10-24-00

WARNING:

Information on this form may become public. Credit card information should not be included on this form. Provide credit card information and authorization on PTO-2038.

Burden Hour Statement This form is estimated to take 0 2 hours to complete Time will vary depending upon the needs of the individual case Any comments on the amount of time you are required to complete this form should be sent to the Chief Information Officer, Patent and Trademark Office, Washington, DC 20231
DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS SEND TO Assistant Commissioner for Patents, Washington, DC 20231

219.39026X00
P10093

UNITED STATES PATENT APPLICATION FOR:

SYSTEM AND METHOD FOR PROVIDING DETAILED PATH INFORMATION TO CLIENTS

Inventor:

RAJESH R. SHAH

Prepared by:

Antonelli, Terry, Stout & Kraus, LLP
1300 North Seventeenth Street, Suite 1800
Arlington, Virginia 22209
Tel: 703/312-6600
Fax: 703/312-6666

TITLE

SYSTEM AND METHOD FOR PROVIDING
DETAILED PATH INFORMATION TO CLIENTS

FIELD

The present invention generally relates to data networks and more particularly relates to a system and method for providing detailed path information to clients.

BACKGROUND

A data network generally includes a network of nodes connected by point-to-point links. Each physical link may support a number of logical point-to-point channels. Each channel may be a bi-directional communication path for allowing commands and message data to flow between two connected nodes within the data network. Each channel may refer to a single point-to-point connection where message data may be transferred between two endpoints or systems. Data may be transmitted in packets including groups called cells from source to destination often through intermediate nodes.

In many data networks, hardware and software may often be used to support asynchronous data transfers between two memory regions, often on different systems. Each system may correspond to a multi-processor system including one or more processors. Each system may serve as a source (initiator) system which initiates a message data transfer (message send operation) or a target system of a message passing operation (message receive operation). Examples of such a multi-processor system may include host servers providing a variety of

applications or services, and I/O units providing storage oriented and network oriented I/O services.

Clients connected on a data network may have multiple ports through which to communicate with other clients or applications on the data network. There are often multiple paths between ports and a large number of ports connected to the network. It is not yet possible for a client to know how many switches and links are traversed from a source to a destination in a particular path in a network. There has been no way to provide clients with information that pertains to the links and switches traversed in various paths.

BRIEF DESCRIPTION OF THE DRAWINGS

A more complete appreciation of example embodiments of the present invention, and many of the attendant advantages of the present invention, will be readily appreciated as the same becomes better understood by reference to the following detailed description when considered in conjunction with the accompanying drawings in which like reference symbols indicate the same or similar components, wherein:

FIG. 1 illustrates an example data network having several nodes interconnected by corresponding links of a basic switch according to an embodiment of the present invention;

FIG. 2 illustrates another example data network having several nodes interconnected by corresponding links of a multi-stage switched fabric according to an embodiment of the present invention;

FIG. 3 illustrates a block diagram of a host system of an example data network according to an embodiment of the present invention;

FIG. 4 illustrates a block diagram of a host system of an example data network according to another embodiment of the present invention;

FIG. 5 illustrates an example software driver stack of a host operating system of an example data network according to an embodiment of the present invention;

FIG. 6 illustrates an example cluster according to an embodiment of the present invention; and

FIG. 7 is a process flow diagram for describing providing detailed path information to clients according to an embodiment of the present invention.

DETAILED DESCRIPTION

Before beginning a detailed description of the subject invention, mention of the following is in order. When appropriate, like reference numerals and characters may be used to designate identical, corresponding or similar components in differing figure drawings. Further, in the detailed description to follow, example sizes/models/values/ranges may be given, although the present invention is not limited to the same.

Clients are distributed throughout a data network. The clients can have multiple ports through which to communicate with other clients and applications in the data network. There are often a large number of paths between ports and a large number of ports connected to the network. Previously, no service has been able to give detailed information to clients in regard to

links and switches traversed in a particular path to a destination. The invention provides detailed information to clients about the links and switches traversed in the available paths from a source to a destination such that clients can make informed decisions about which paths they should use when multiple paths are available.

The present invention is applicable for use with all types of computer networks, I/O hardware adapters and chipsets that allow multiple addresses to be specified to a port, including follow-on chip designs which link together end stations such as computers, servers, peripherals, storage devices, and communication devices for data communications.

Attention now is directed to the drawings and particularly to FIG. 1, in which a simple data network 10 having several interconnected nodes for data communications according to an embodiment of the present invention is illustrated. As shown in FIG. 1, the data network 10 may include, for example, one or more centralized switches 100 and four different nodes A, B, C, and D. Each node (endpoint) may correspond to one or more I/O units and host systems including computers and/or servers on which a variety of applications or services are provided. Each I/O unit may include one or more I/O controllers connected thereto. Each I/O controller may operate to control one or more I/O devices, such as storage devices (e.g., a hard disk drive or tape drive) locally or remotely via a local area network (LAN) or a wide area network (WAN), for example.

The centralized switch 100 may contain, for example, switch ports 0, 1, 2, and 3 each connected to a corresponding node of the four different nodes A, B, C, and D via a corresponding physical link 110, 112, 114, and 116. Each physical link may support a number of logical point-to-point channels. Each channel may be a bi-directional communication path for allowing

commands and data to flow between two connected nodes (e.g., host systems, switch/switch elements, and I/O units) within the network.

Each channel may refer to a single point-to-point connection where data may be transferred between endpoints (e.g., host systems and I/O units). The centralized switch 100 may also contain routing information using, for example, explicit routing and/or destination address routing for routing data from a source node (data transmitter) to a target node (data receiver) via corresponding link(s), and re-routing information for redundancy.

The specific number and configuration of end stations (e.g., host systems and I/O units), switches and links shown in FIG. 1 is provided simply as an example data network. A wide variety of implementations and arrangements of a number of end stations (e.g., host systems and I/O units), switches and links in all types of data networks may be possible.

According to an example embodiment or implementation, the end stations (e.g., host systems and I/O units) of the example data network shown in FIG. 1 may be compatible with the "Next Generation Input/Output (NGIO) Specification" as set forth by the NGIO Forum on July 20, 1999. According to the NGIO Specification, the switch 100 may be an NGIO switched fabric (e.g., collection of links, switches and/or switch elements connecting a number of host systems and I/O units), and the endpoint may be a host system including one or more host channel adapters (HCAs), or a target system such as an I/O unit including one or more target channel adapters (TCAs). Both the host channel adapter (HCA) and the target channel adapter (TCA) may be broadly considered as fabric adapters provided to interface endpoints to the NGIO switched fabric, and may be implemented in compliance with "Next Generation I/O Link

Architecture Specification: HCA Specification, Revision 1.0" as set forth by NGIO Forum on May 13, 1999 for enabling the endpoints (nodes) to communicate to each other over an NGIO channel(s).

For example, FIG. 2 illustrates an example data network 10' using an NGIO architecture to transfer data from a source node to a destination node according to an embodiment of the present invention. As shown in FIG. 2, the data network 10' includes an NGIO fabric 100' (multi-stage switched fabric comprised of a plurality of switches) for allowing a host system and a remote system to communicate to a large number of other host systems and remote systems over one or more designated channels. A single channel may be sufficient but data transfer spread between adjacent ports can decrease latency and increase bandwidth. Therefore, separate channels for separate control flow and data flow may be desired. For example, one channel may be created for sending request and reply messages. A separate channel or set of channels may be created for moving data between the host system and any ones of target systems. In addition, any number of end stations, switches and links may be used for relaying data in groups of cells between the end stations and switches via corresponding NGIO links.

For example, node A may represent a host system 130 such as a host computer or a host server on which a variety of applications or services are provided. Similarly, node B may represent another network 150, including, but not limited to, local area network (LAN), wide area network (WAN), Ethernet, ATM and fiber channel network, that is connected via high speed serial links. Node C may represent an I/O unit 170, including one or more I/O controllers and I/O units connected thereto. Likewise, node D may represent a remote system 190 such as a

target computer or a target server on which a variety of applications or services are provided. Alternatively, nodes A, B, C, and D may also represent individual switches of the multi-stage switched fabric 100' which serve as intermediate nodes between the host system 130 and the remote systems 150, 170 and 190.

The multi-state switched fabric 100' may include a central network manager 250 connected to all the switches for managing all network management functions. However, the central network manager 250 may alternatively be incorporated as part of either the host system 130, the second network 150, the I/O unit 170, or the remote system 190 for managing all network management functions. In either situation, the central network manager 250 may be configured for learning network topology, determining the switch table or forwarding database, detecting and managing faults or link failures in the network and performing other network management functions.

A host channel adapter (HCA) 120 may be used to provide an interface between a memory controller (not shown) of the local system 130 and a multi-stage switched fabric 100' via high speed serial NGIO links. Similarly, target channel adapters (TCA) 140 and 160 may be used to provide an interface between the multi-stage switched fabric 100' and an I/O controller of either a second network 150 or an I/O unit 170 via high speed serial NGIO links. Separately, another target channel adapter (TCA) 180 may be used to provide an interface between a memory controller (not shown) of the remote system 190 and the multi-stage switched fabric 100' via high speed serial NGIO links. Both the host channel adapter (HCA) and the target channel adapter (TCA) may be broadly considered as fabric hardware adapters provided to interface either the

host system 130 or any one of the target systems 150, 170 and 190 to the switched fabric, and may be implemented in compliance with "Next Generation I/O Link Architecture Specification: HCA Specification, Revision 1.0" as set forth by NGIO Forum on May 13, 1999 for enabling the endpoints (nodes) to communicate to each other over an NGIO channel(s). However, NGIO is merely one example embodiment or implementation of the present invention, and the invention is not limited thereto. Rather, the present invention may be applicable to a wide variety of any number of data networks, hosts and I/O units. For example, practice of the invention may also be made with Future Input/Output (FIO) and/or InfiniBand technologies. FIO specifications have not yet been released, owing to subsequent agreement of NGIO and FIO factions to combine efforts on InfiniBand. InfiniBand information/specifications are presently under development and will be published in a document entitled "InfiniBand Architecture Specification" by the InfiniBand Trade Association (formed August 27, 1999) having the Internet address of "http://www.InfiniBandta.org". The "InfiniBand Architecture Specification" describes features and benefits which are complementary to those provided by NGIO and FIO technologies, and are similarly useful.

Returning to discussions, one example embodiment of a host system 130 is shown in FIG. 3. Referring to FIG. 3, the host system 130 may correspond to a multi-processor system, including one or more processors 202A-202N coupled to a host bus 203. Each of the multiple processors 202A-202N may operate on a single item (I/O operation), and all of the multiple processors 202A-202N may operate on multiple items (I/O operations) on a list at the same time. An I/O and memory controller 204 (or chipset) may be connected to the host bus 203. A main

memory 206 may be connected to the I/O and memory controller 204. An I/O bridge 208 may operate to bridge or interface between the I/O and memory controller 204 and an I/O bus 205. Several I/O controllers may be attached to the I/O bus 205, including I/O controllers 210 and 212. I/O controllers 210 and 212 (including any I/O devices connected thereto) may provide bus-based I/O resources.

One or more host-fabric adapters 120 may also be connected to the I/O bus 205. Alternatively, one or more host-fabric adapters 120 may be connected directly to the I/O and memory controller (or chipset) 204 to avoid the inherent limitations of the I/O bus 205 as shown in FIG. 4. In either embodiment, one or more host-fabric adapters 120 may be provided to interface the host system 130 to the multi-stage switched fabric 100'.

FIGS. 3-4 merely illustrate example embodiments of a host system 130. A wide array of processor configurations of such a host system 130 may be available. Software driver stack for the host-fabric adapter 120 may also be provided to allow the host system 130 to exchange data with one or more remote systems 150, 170 and 190 via the switched fabric 100', while preferably being compatible with many currently available operating systems, such as Windows 2000.

FIG. 5 illustrates an example software driver stack of a host system 130. As shown in FIG. 5, a host operating system (OS) 500 may include a kernel 510, an I/O manager 520, and a plurality of channel drivers 530A-530N for providing an interface to various I/O controllers. Such a host operating system (OS) 500 may be Windows 2000, for example, and the I/O manager 520 may be a Plug-n-Play manager.

In addition, a host-fabric adapter software stack (driver module) may be provided to access the switched fabric 100' and information about fabric configuration, fabric topology and connection information. Such a host-fabric adapter software stack (driver module) may include a fabric bus driver 540 and a fabric adapter device-specific driver 550 utilized to establish communication with a remote fabric-attached agent (e.g., I/O controller), and perform functions common to most drivers, including, for example, host-fabric adapter initialization and configuration, channel configuration, channel abstraction, resource management, fabric management service and operations, send/receive I/O transaction messages, remote direct memory access (RDMA) transactions (e.g., read and write operations), queue management, memory registration, descriptor management, message flow control, and transient error handling and recovery. Such software driver module may be written using high-level programming languages such as C, C++ and Visual Basic, and may be provided on a computer tangible medium, such as memory devices; magnetic disks (fixed, floppy, and removable); other magnetic media such as magnetic tapes; optical media such as CD-ROM disks, or via Internet downloads, which may be available for a fabric administrator to conveniently plug-in or download into an existing operating system (OS). Such a software driver module may also be bundled with the existing operating system (OS) which may be activated by a particular device driver.

The host-fabric adapter driver module may consist of three functional layers: a HCA services layer (HSL), a HCA abstraction layer (HCAAL), and a HCA device-specific driver (HDSD) in compliance with the "Next Generation I/O Architecture: Host Channel Adapter Software Specification." For example, the HCA service layer (HSL) may be inherent to all

channel drivers 530A-530N for providing a set of common fabric services in a service library, including connection services, resource services, and HCA services required by the channel drivers 530A-530N to instantiate and use NGIO channels for performing data transfers over the NGIO channels.

The host system 130 may also communicate with one or more remote systems 150, 170 and 190, including I/O units and I/O controllers (and attached I/O devices) which are directly attached to the switched fabric 100' (i.e., the fabric-attached I/O controllers) using a Virtual Interface (VI) architecture in compliance with the "Virtual Interface (VI) Architecture Specification, Version 1.0," as set forth by Compaq Corp., Intel Corp., and Microsoft Corp., on December 16, 1997. NGIO and VI architectures support asynchronous data transfers between two memory regions, typically on different systems over one or more designated channels of a data network. Each system using a VI architecture may contain work queues formed in pairs including a send queue and a receive queue in which requests, in the form of descriptors, are posted to describe data movement operation and location of data to be moved for processing and/or transportation via a NGIO switched fabric. The VI Specification defines VI mechanisms for low-latency, high-bandwidth message-passing between interconnected nodes connected by multiple logical point-to-point channels. Other architectures such as InfiniBand may also be used to implement the present invention.

In such a data network, NGIO, VI and InfiniBand hardware and software may be used to support asynchronous data transfers between two memory regions, often on different systems. Each system may serve as a source (initiator) system which initiates a message data transfer

(message send operation) or a target system of a message passing operation (message receive operation). Each system may correspond to a multi-processor system including multiple processors each capable of processing an I/O completion on a different shared resource (such as work queues or other memory elements associated with a given hardware adapter). Examples of such a multi-processor system may include host servers providing a variety of applications or services, and I/O units providing storage-oriented and network-oriented I/O services.

The InfiniBand architecture specification defines a common base for connecting hosts and I/O enclosures together in a cluster for improved performance. A cluster that conforms to the InfiniBand architecture specification allows hardware and software solutions from different vendors to inter-operate, and is often referred to as a "subnet". Moreover, a cluster may include one or more subnets.

A group of hosts and I/O enclosures in an InfiniBand cluster is managed by a subnet manager. One of the hosts can be designated the subnet manager. Each host system or I/O enclosure is connected to the interconnection fabric through a channel adapter. A channel adapter may have one or more connection points called ports.

The subnet manager assigns each port at least one unique address denoted a "local identification value" (LID). The subnet manager operates to discover fabric topology, assign unique addresses to all channel adapter ports that are connected to the fabric, program switch forwarding tables, and prepare all fabric connected agents so that they can communicate with other fabric agents, in addition to performing other tasks.

According to the InfiniBand architecture specification, multiple LIDs can be assigned to each port. Each LID assigned to a port represents a unique path to this port from some other port on the cluster (or subnet). A client that wants to use multiple paths to a remote client can use different LIDs to specify different paths to its destination through the fabric. This allows a client to perform load balancing, obtain better throughput, and recover from the failure of one path if some alternate path is still functional.

To enable multi-pathing, the subnet manager identifies all possible paths to a port from any other port on the fabric, and then assigns enough LIDs to the port such that different paths to this port can be identified by a different LID. If multiple paths exist between two ports, clients on each port can explicitly choose which path will be used based on the LIDs used to communicate between the two ports. Multiple paths will exist if the subnet contains multiple (redundant) links that connect switches or channel adapters together.

In today's high-performance computing environment, clusters are becoming more popular because of the better properties they exhibit compared to individual high-performance servers and workstations. A cluster includes one or more host nodes and zero or more I/O enclosures connected together by a (typically high-speed) interconnection fabric. Clusters are typically based on a unifying technology that makes it easier to plug in solutions from different vendors. Examples of such technologies are InfiniBand and Fiber Channel.

Some of the important benefits of clustering based on InfiniBand technology are the ability to support high bandwidth, virtually unlimited scalability, and good fault isolation characteristics. A client, such as a host or an I/O enclosure, that is aware that it is running on an

InfiniBand cluster can take advantage of multiple paths between that client and another client with which it is trying to communicate. For example, a pair of InfiniBand clients might decide to use a primary path for communication and fail-over to an alternate path if the primary path fails. Fail-over is the process of using a new path between a pair of fabric attached agents/clients when an existing path breaks. Clients at both ends of the path need to fail over to the new path. Another pair of InfiniBand clients might want to simultaneously use multiple paths for higher throughput or load balancing.

FIG. 6 shows an example cluster (i.e., a subnet) in which multiple paths exists between various clients. The cluster shown in FIG. 6 includes three interconnected switches, a first switch S1, a second switch S2, and a third switch S3. The cluster includes a first host 602, a second host 604, and a third host 606. The second host 604 serves as the subnet manager in the example cluster shown in FIG. 6. The cluster includes an I/O enclosure 608.

In the example, the first host 602 includes a first channel adapter 610. The first channel adapter 610 presents a first port P1 and a second port P2. The second host 604 includes a second channel adapter 612. The second channel adapter 612 presents a third port P3 and a fourth port P4. The third host 606 includes a third channel adapter 614. The third channel adapter 614 presents a fifth port P5. The I/O enclosure 608 includes a fourth channel adapter 616. A first I/O controller 618 and a second I/O controller 620 are coupled to the fourth channel adapter 616. The fourth channel adapter 616 presents a sixth port P6 and a seventh port P7.

A set of links provides a communicative capability for the cluster shown in FIG. 6. A first link L1 connects the first port P1 and the first switch S1. A second link L2 connects the

second port P2 and the second switch S2. A third link L3 connects the sixth port P6 and the first switch S1. A fourth link L4 connects the first switch S1 and the third switch S3. A fifth link L5 connects the first switch S1 and the second switch S2. A sixth link L6 connects the second switch S2 and the third port P3. A seventh link L7 connects the second switch S2 and the third switch S3. An eighth link L8 connects the third switch S3 and the fourth port P4. A ninth link L9 connects the third switch S3 and the seventh port P7. A tenth link L10 connects the third switch S3 and the fifth port P5.

In the example cluster depicted in FIG. 6, the ports are numbered from P1 to P7, the switches are numbered from S1 to S3, and the links are numbered from L1 to L10. The physical topology of the subnet is such that the following paths are available to an InfiniBand client running on the first host 602 that wants to communicate with an InfiniBand client running on the second host 604:

Path One: from port P2 to link L2 to switch S2 to link L6 to port P3. This path traverses two links and one switch.

Path Two: from port P2 to link L2 to switch S2 to link L7 to switch S3 to link L8 to port P4. This path traverses three links and two switches.

Path Three: from port P1 to link L1 to switch S1 to link L5 to switch S2 to link L6 to port P3. This path traverses three links and two switches.

Path Four: from port P1 to link L1 to switch S1 to link L4 to switch S3 to link L8 to port P4. This path traverses three links and two switches.

Path Five: from port P2 to link L2 to switch S2 to link L5 to switch S1 to link L4 to switch S3 to link L8 to port P4. This path traverses four links and three switches.

Path Six: from port P1 to link L1 to switch S1 to link L4 to switch S3 to link L7 to switch S2 to link L6 to port P3. This path traverses four links and three switches.

Path Seven: from port P1 to link L1 to switch S1 to link L5 to switch S2 to link L7 to switch S3 to link L8 to port P4. This path traverses four links and three switches.

Even in a small subnet, several paths may be available between a pair of clients. The quality of the available paths can vary widely. The metric that is used to evaluate the quality of a path can be different based on the reason why the path is being used.

For example, if multiple paths are being used for fail-over, an important metric that determines the quality of the available paths is whether the paths traverse common links or switches. If the primary and alternate paths traverse a large number of common switches or links, a failure in one of those switches or links will break not just the primary path but also the alternate path. As the link/switch overlap between the primary and alternate path increases, the probability that both paths will fail simultaneously increases. For fail-over, a client pair may want to use paths that have as few overlapping switches and links as the physical topology allows, even though these paths do not have the best path-latency or hop-count properties. The InfiniBand architecture specification does not define or provide mechanisms for reporting such detailed path information to interested clients.

In the example cluster illustrated in FIG. 6, only Path One and Path Four have no common link or switch. All other path-pairs share at least one common link or switch. For this

reason, a client running on the first host 602 would want to chose Path One as the primary path and Path Four as the alternate path for fail-over in order to communicate with an InfiniBand client running on the second host 604.

The only way it can make this informed decision is if it has detailed information about these paths. The hop count value is not sufficient to make this decision since the hop count is the same for Path Two, Path Three, and Path Four (two switches and three links traversed). In the absence of detailed path information, a client may be forced to blindly cycle through multiple available paths until the client encounters a functional alternate path when the primary path fails. This may consume a large amount of time.

The situation is made worse by the fact that unreliable InfiniBand datagrams may be used to verify the availability and to set up a connection using the alternate path when the primary path fails. An InfiniBand datagram is a fixed-size message that is used to communicate between fabric-attached end points. A Queue Pair (QP) that is configured for sending or receiving datagram messages can simultaneously send and receive from multiple InfiniBand end points.

The InfiniBand client failing over may have to wait for a large time-out period before such client can decide whether an alternate path it is trying to use is also broken or just temporarily unavailable for some other reason. This makes it very difficult to implement fail-over since the extended delays in failing over to a functional path may trigger timeouts in the consumers of the InfiniBand services.

For example, if a host-side I/O controller driver for a fabric-attached SCSI adapter wants to use multiple paths for fail-over, it must fail-over to a new path within a few seconds when the

primary path fails. If it is unable to do so, the upper-level SCSI drivers will time out and attempt to reset the SCSI bus.

The need for detailed path information also exists for a client pair that wants to use multiple paths to facilitate load distribution or higher throughput. Such client pair might want to balance the importance the client pair provides to other path properties (like link speed, service levels supported, path latency, and hop count) versus the number of overlapping switches or links. Even though there may not be a single correct answer in regard to the available paths to use, clients ought to be given enough detailed information about the available paths so that they can make an informed decision that is appropriate for them.

The InfiniBand architecture specification defines a path record that provides some properties of a path. Properties reported in the path record include the hop count, the service levels supported, the maximum transfer unit (MTU), link speed and latency cost of the path. There is, however, no capability of providing detailed link or switch traversal information to interested clients and no mechanisms are provided to query or report this information.

The invention provides a procedure that can be used by interested InfiniBand clients to obtain detailed path-composition information, which allows clients to make informed decisions about which paths should be used to best suit their purposes. According to the invention, an InfiniBand subnet has a service provider that provides detailed information about which links and switches are traversed in a path. This allows clients to make informed decisions about which paths to use when multiple paths are available. Two separate ways in which such a service can be implemented are described as follows.

First, the detailed path information service can be implemented as a service agent sitting on top of the General Service Interface (GSI). The General Service Interface (GSI) is an interface providing management services (e.g., connection, performance, and diagnostics) other than subnet management. Queue Pair 1 (QP1) is reserved for the GSI, which may redirect requests to other Queue Pairs (QPs).

Queries and responses to and from this service are sent using management datagrams (MADs) sent on queue pair 1. A Management Datagram (MAD) refers to the contents of an unreliable datagram packet used for communication among the HCAs, switches, routers, and TCAs to manage the network. The InfiniBand architecture specification describes the format of a number of these management commands.

The service that provides detailed path information registers with the GSI as a service agent. One option is that this service agent is implemented by the subnet administration code that also responds to the SubnAdm class of MADs. This is a natural fit since the subnet administrator is also responsible for providing other path information (like path latency, hop count, service classes supported, the maximum transfer unit and path speed) as described in the InfiniBand architecture specification. Since the MAD format for querying and reporting detailed path information is not defined in the InfiniBand architecture specification, vendor-specific MADs can be used for this purpose.

The general header format of a vendor-specific MAD is defined in the InfiniBand architecture specification. To issue the path information query, a client would send a message with class value set to VendorSpecific; method value set to VendorSpecificGet or

VendorSpecificGetTable; and attribute value set to DetailedPathInfo. This message would be sent to the subnet administrator address. If the service resides at a different local identification value (LID) or queue pair, the client can be redirected using the ClassPortInfo message specified in the InfiniBand architecture specification. As input, the client would supply relevant information like the LID or Global Identifier (GID) of the source and destination. A GID is a 128-bit identifier used to identify a port on a channel adapter, a port on a router, or a multicast group. A GID is a valid 128-bit IPv6 address (per RFC 2373) with additional properties or restrictions defined within the InfiniBand architecture specification to facilitate efficient discovery, communication, and routing.

Different implementations can also take the node GUID (Globally Unique Identifier) or platform GUID of the source and destination as input. A GUID is a software-readable number that uniquely identifies a device or component. As output of this query, the subnet administrator provides the port and node GUIDs of all switches that are traversed in this path. The width and layout of the input and output fields in the MAD are specified and documented by whoever implements the service. Note that links themselves do not have any identification or visibility and cannot be directly listed in the path information. However, the port GUIDs of the switch ports listed in the path information will uniquely identify the links being traversed. The switch port GUIDs and node GUIDs are listed in the order they are traversed from the source to the destination. For some queries, multiple packets may be needed to report the results. In this case, the mechanisms that are used to send multi-packet responses for other SubnAdm messages can be used here also.

The advantage of the foregoing arrangement is that the infrastructure in place to query and report other path properties can be used with only minor modifications to query and report detailed path information. Redirection to a different LID or queue pair can be accomplished using the infrastructure already put in place to redirect other service classes.

FIG. 7 illustrates a path service implemented as a vendor-specific service over the GSI and the process performed by the path service agent. Regarding FIG. 7, in block 702, an interested client sends a VendorSpecificGet(DetailedPathInfo) message to the service, where the input values in the message identify the path for which details are requested. In block 704, the service determines whether the request needs to be redirected. If yes, in block 706, the service returns a ClassPortInfo response to the client with relevant information, by which the client is instructed to submit the request to the redirected address. If no, the request does not need to be redirected, in block 708, the service sends a response containing the port GUIDs and node GUIDs of the one or more switches traversed in the path, in sequential order from source to destination. For a multi-packet response, the service uses the same mechanism for the multi-packet response as for the SubnAdm class MADs.

Second, the detailed path information service can also be implemented as a service that uses regular (i.e. non-MAD) unreliable datagrams to communicate with clients. Clients use the service ID resolution protocol defined in the InfiniBand architecture specification to communicate with this service. A client wishing to query this service first needs to query a service locator that provides the address (LID) where the service resides. The client then sends a service ID resolution request (SIDR_REQ) message to this address. As a response, the client

receives a service ID resolution response (SIDR_REP) message that provided other information (like queue pair and Q-Key) needed to be able to communicate with this service. Once the client has all the information needed to communicate with the path service, the client sends a query to the service. As input, the client supplies relevant information like the LID or GID of the source and destination.

Different implementations can also take the node GUID or platform GUID of the source and destination as input. As output of this query, the detailed path information service provides the port GUIDs and node GUIDs of all switches that are traversed in this path. The format and layout of the input and output fields in the message are specified and documented by whoever implements the path service. The switch port GUIDs and node GUIDs are listed in the order they are traversed from the source to the destination.

For some queries, multiple packets may be needed to report the results. In this case, the service implementation defines the mechanisms that are used to send multi-packet responses. The advantage of this implementation is that a vendor-specific MAD is not needed for communication. This can be an important issue since the InfiniBand architecture specification allows only one type of vendor-specific MAD to be used in a subnet. If two separate vendors want to use vendor specific MADs in the same subnet, they cannot do that without explicitly cooperating with each other.

The ability to use multiple paths to the same destination is an important benefit of clusters over traditional bus-based systems. The invention makes available information that is valuable to intelligent clients that want to benefit from multiple paths to their destination. Such clients

can make an informed decision about which of the several available paths they should use based on this detailed path information.

While there have been illustrated and described what are considered to be example embodiments of the present invention, it will be understood by those skilled in the art and as technology develops that various changes and modifications may be made, and equivalents may be substituted for elements thereof without departing from the true scope of the present invention. For example, the present invention is applicable to all types of data networks that allow multiple addresses to be assigned to ports, including, but not limited to, a local area network (LAN), a wide area network (WAN), a campus area network (CAN), a metropolitan area network (MAN), a global area network (GAN) and a system area network (SAN). Further, many other modifications may be made to adapt the teachings herein to a particular situation without departing from the scope thereof. Therefore, it is intended that the present invention not be limited to the various example embodiments disclosed, but that the present invention includes all embodiments falling within the scope of the appended claims.

WHAT IS CLAIMED IS:

1. A method, comprising:

performing a topology discovery of a cluster that includes a plurality of ports;

identifying all the possible paths to each port from any other port;

receiving a request from a client, the request identifying a source and a destination of a path; and

sending a response to the client based on the request, the response identifying one or more links and switches between the source and the destination.

2. The method of claim 1, further comprising:

determining whether the request ought to be redirected.

3. The method of claim 2, further comprising:

sending a redirection address to the client if the request ought to be redirected.

4. The method of claim 3, further comprising:

submitting the request to the redirection address.

5. The method of claim 1, wherein:

the source and the path are each identified by a respective local identification value (LID).

6. The method of claim 1, wherein:

each switch is identified by a respective globally unique identifier (GUID), and each link is identified by the port GUIDs of ports connected to both ends of the link.

7. The method of claim 1, wherein:

the response identifies an order in which the one or more links and switches are traversed from the source to the destination.

8. A cluster, comprising:

a fabric of switches;

a plurality of ports on the fabric;

a service coupled to the fabric;

wherein the service is operative to send a response based on a request from a client;

wherein the request identifies a source and a destination of a path; and

wherein the response identifies one or more links and switches between the source and the destination.

9. The cluster of claim 8, wherein:

the client is a host.

10. The cluster of claim 8, wherein:

the client is an I/O enclosure.

11. The cluster of claim 8, wherein:

the service is operative to determine whether the request ought to be redirected.

12. The cluster of claim 11, wherein:

if the request ought to be redirected, the service is operative to send a redirection address to the client.

13. The cluster of claim 12, wherein:

the request is submitted to the redirection address.

14. The cluster of claim 8, wherein:

the response identifies an order in which the one or more links and switches are traversed from the source to the destination.

15. The cluster of claim 8, wherein:

the service is operative to identify all the possible paths to each port from any other port.

16. A computer readable medium having stored thereon instructions which, when executed by a processor, cause the processor to perform a method, said method comprising:

performing a topology discovery of a cluster that includes a plurality of ports;

identifying all the possible paths to each port from any other port;

receiving a request from a client, the request identifying a source and a destination of a path; and

sending a response to the client based on the request, the response identifying one or more links and switches between the source and the destination.

17. The computer readable medium of claim 16, said method further comprising:

determining whether the request ought to be redirected.

18. The computer readable medium of claim 17, said method further comprising:

sending a redirection address to the client if the request ought to be redirected.

19. The computer readable medium of claim 18, said method further comprising:

submitting the request to the redirection address.

20. The computer readable medium of claim 16, wherein:

the source and the path are each identified by a respective local identification value (LID).

21. The computer readable medium of claim 16, wherein:

each switch is identified by a respective globally unique identifier (GUID), and each link is identified by the port GUIDs of ports connected to both ends of the link.

22. The computer readable medium of claim 16, wherein:

the response identifies an order in which the one or more switches are traversed from the source to the destination.

ABSTRACT

Clients are able to use a service to obtain detailed path information. The service provides detailed information concerning which links and switches are traversed in a path in a cluster. The service allows clients to make informed decisions about which paths to use when multiple paths are available. The cluster includes a fabric of switches and a plurality of ports. A client makes a request to the service. The request identifies a source and a destination of a path. The service is operative to send a response to the request. The response identifies one or more links and switches between the source and the destination.

FIG. 1

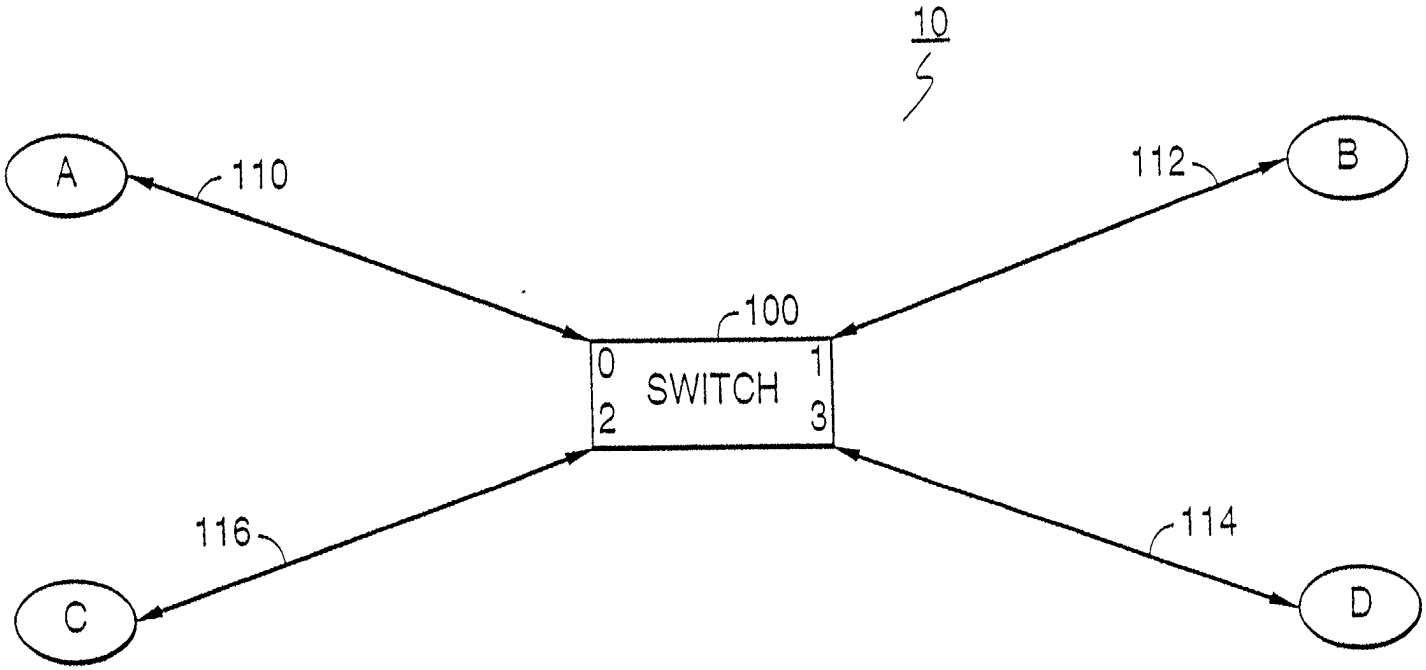
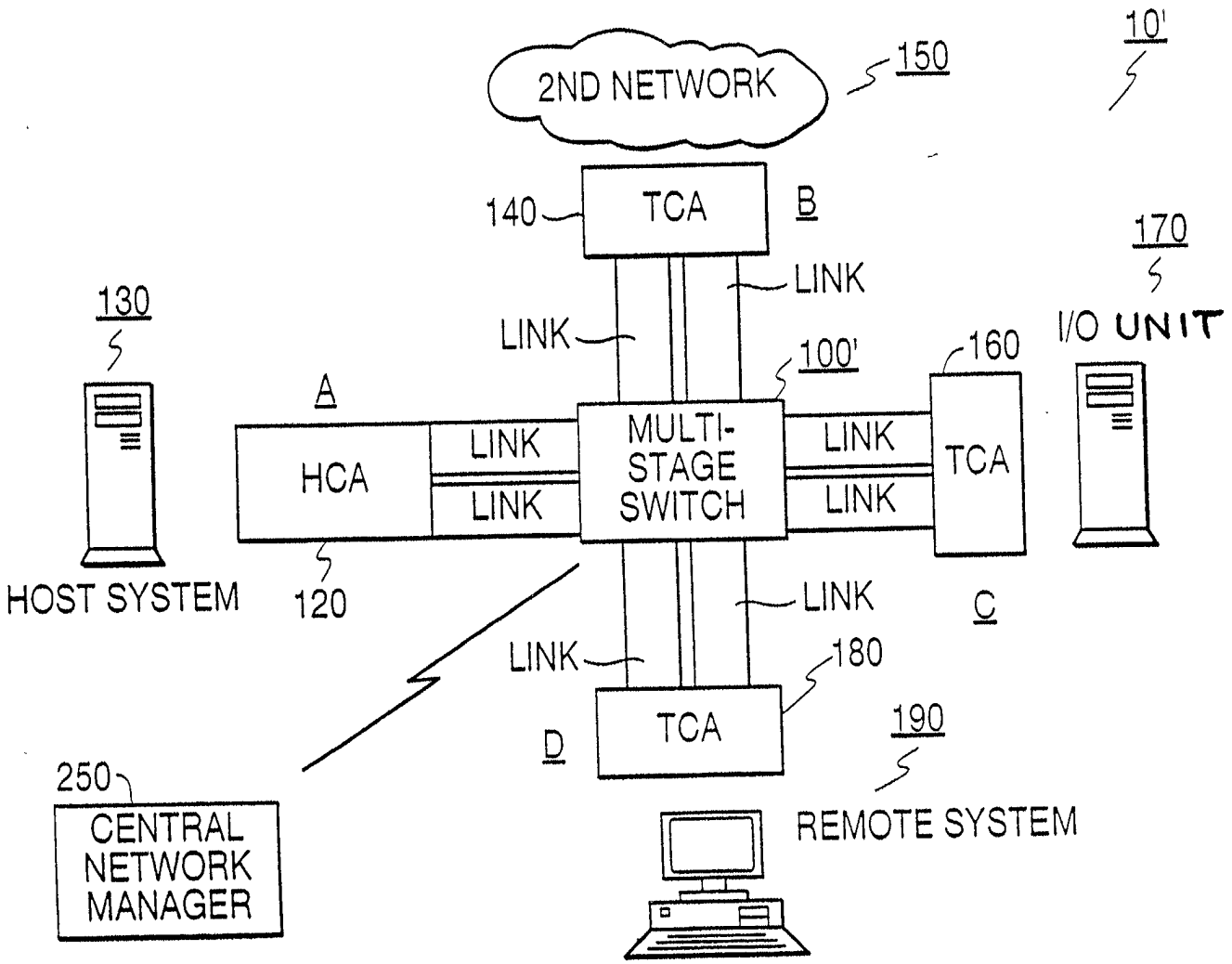


FIG. 2



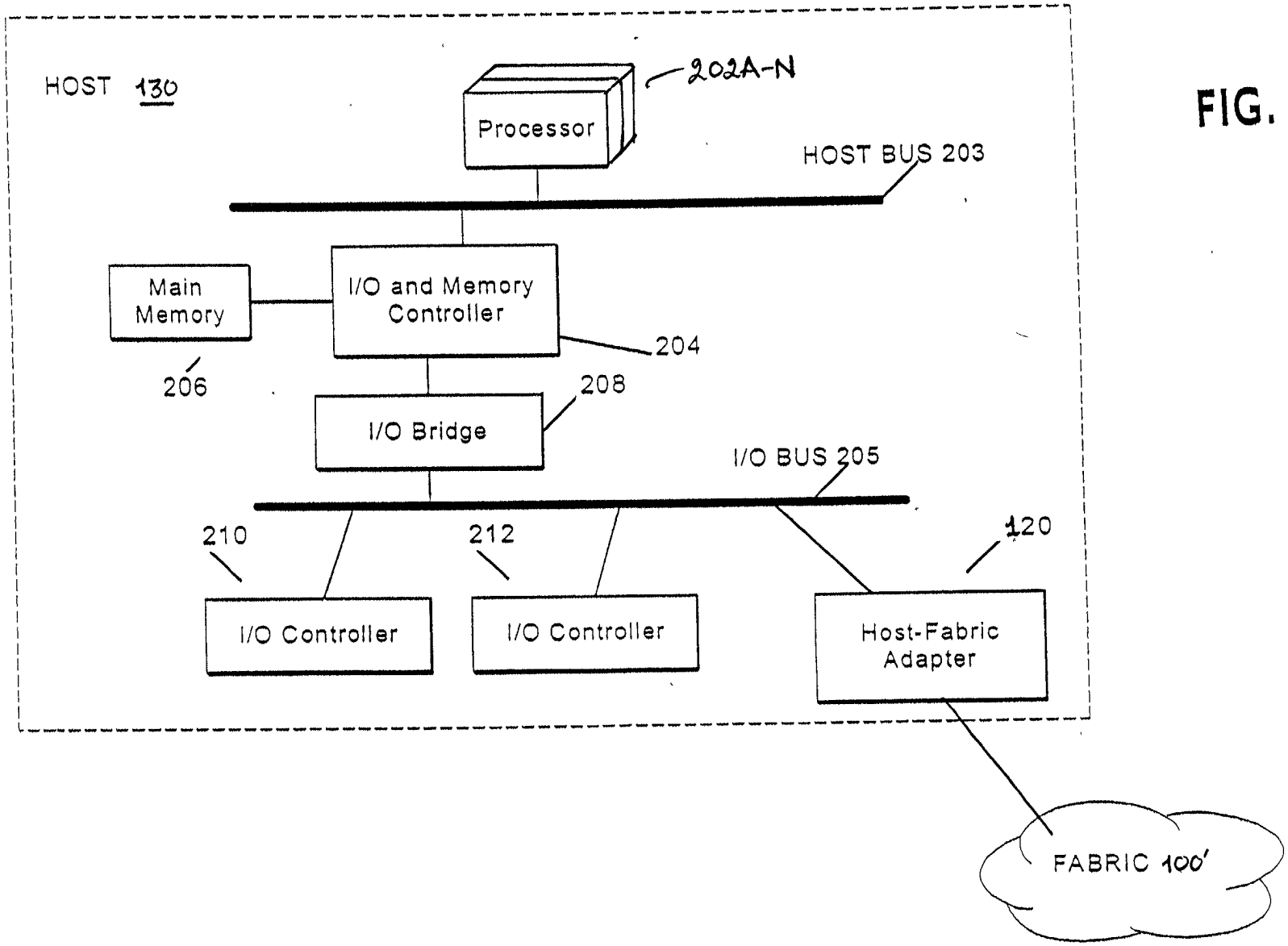


FIG. 3

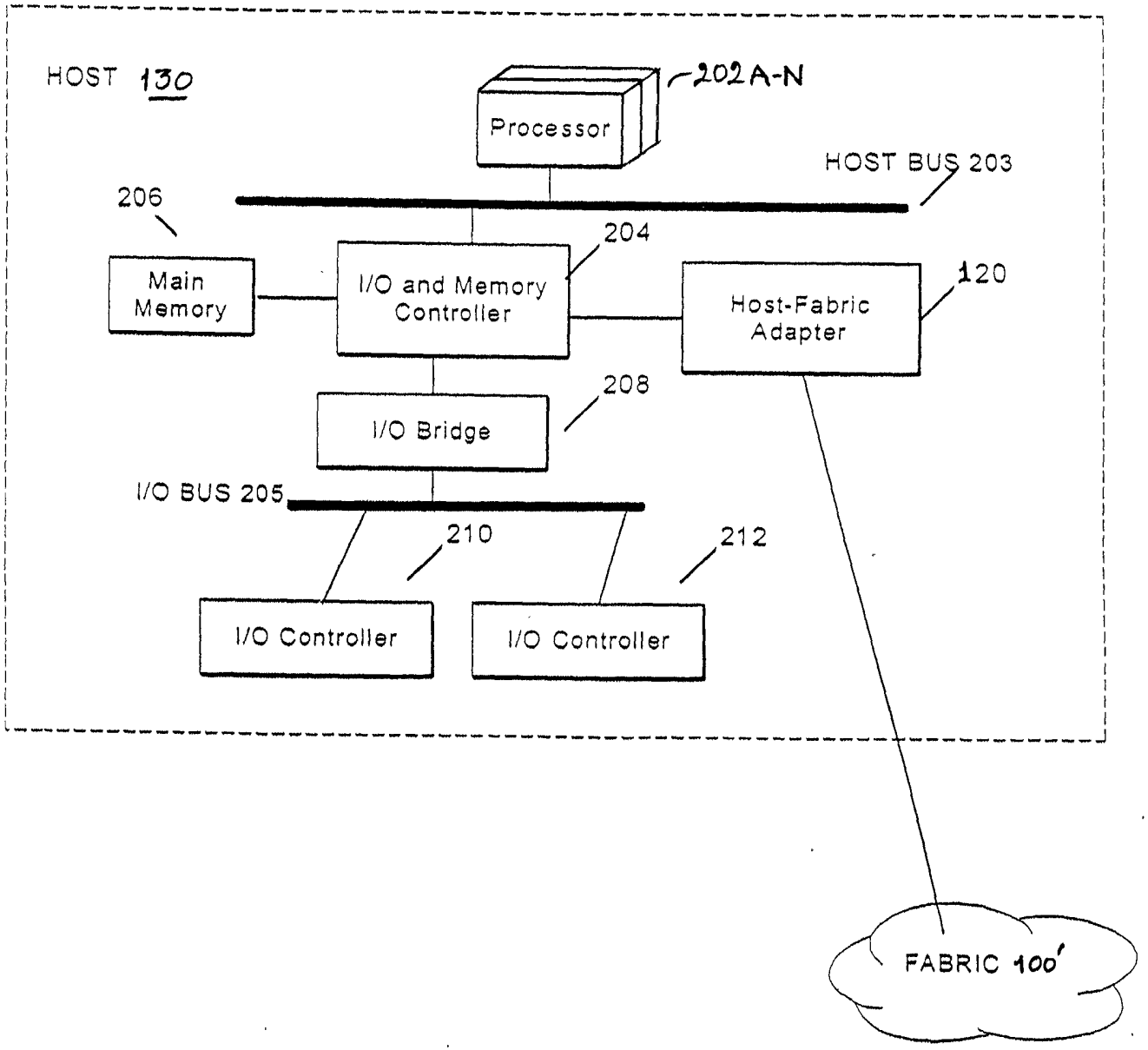
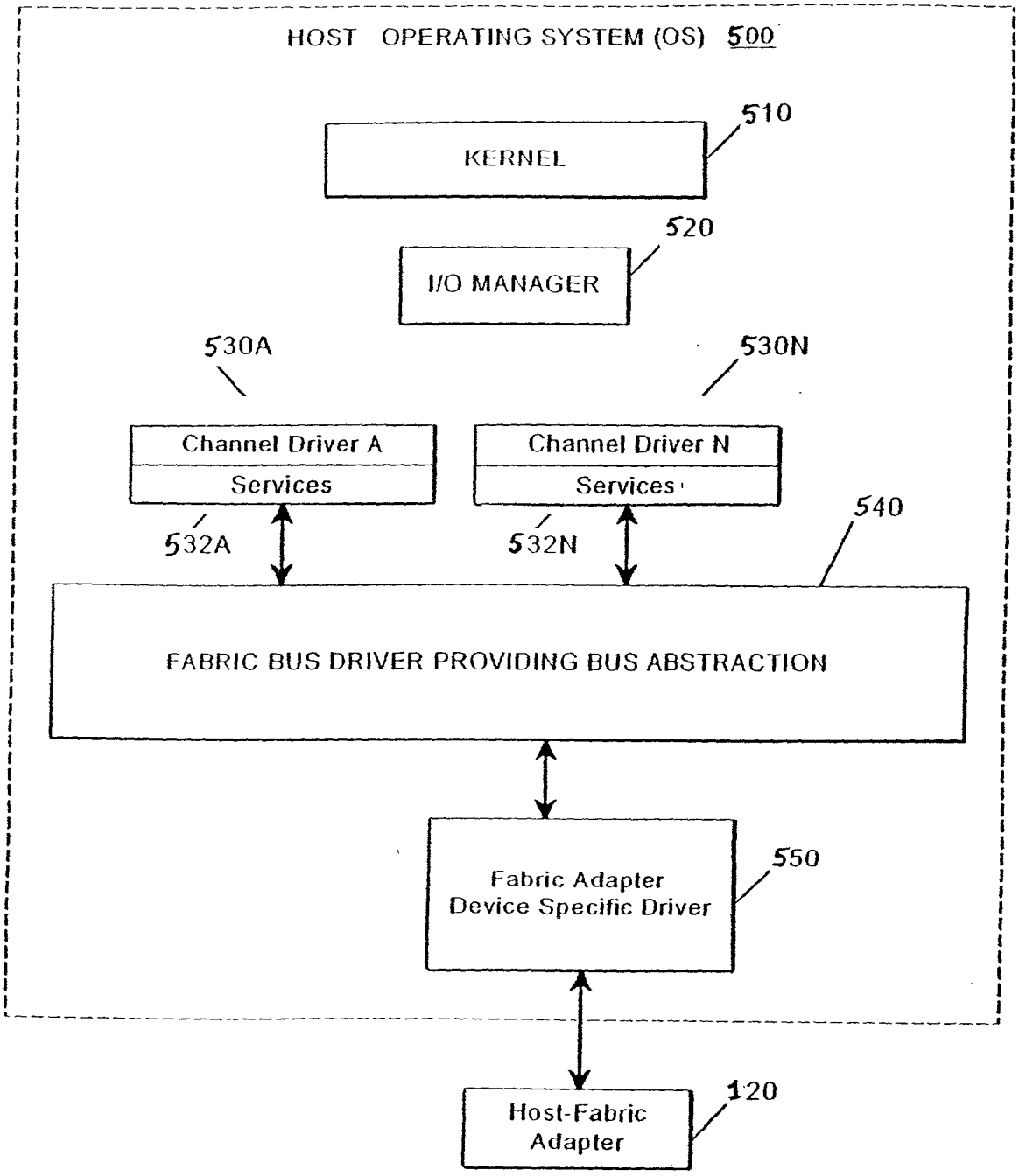

HOST 130
Processor
202A-N
HOST BUS 203
206
204
Main Memory
I/O and Memory Controller
120
Host-Fabric Adapter
208
I/O Bridge
I/O BUS 205
210
212
I/O Controller
I/O Controller
FABRIC 100'

FIG. 4



EXAMPLE SOFTWARE DRIVER STACKS FOR HOST SYSTEM

FIG. 5

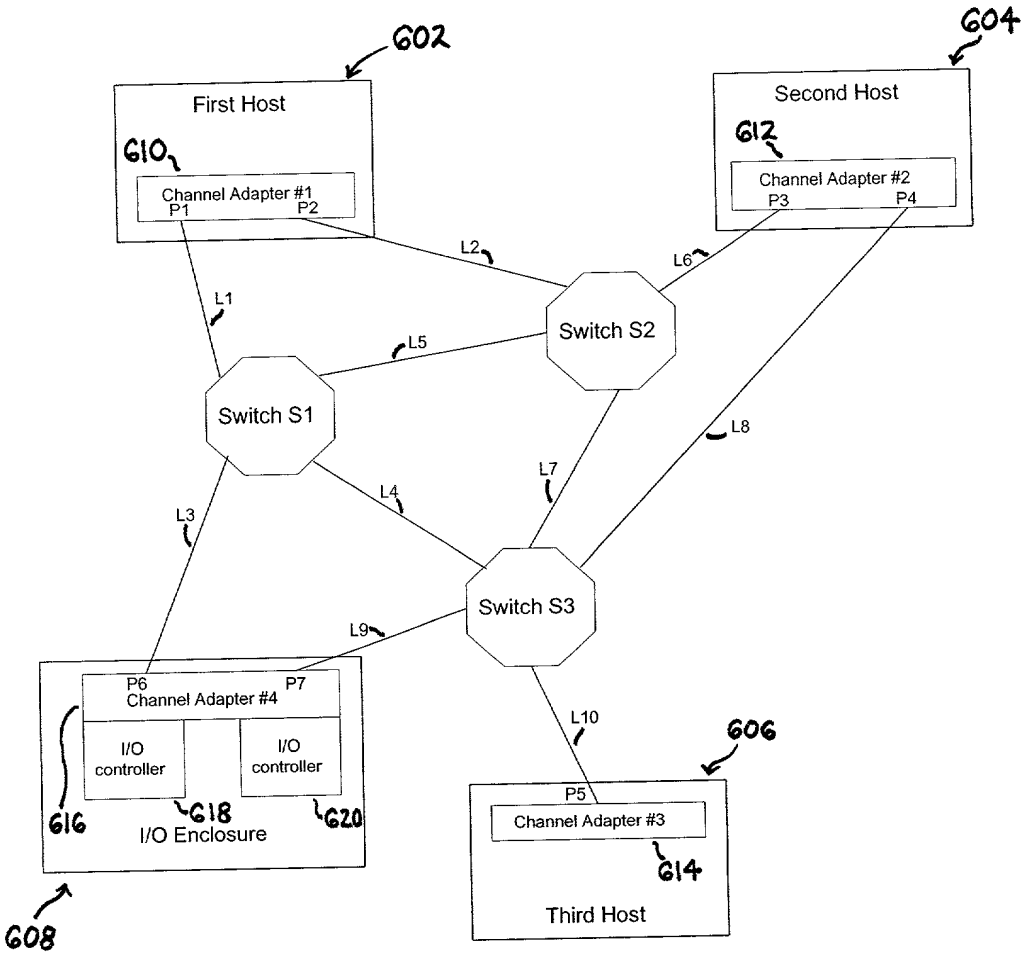

602
First Host
610
Channel Adapter #1
P1
P2
604
Second Host
612
Channel Adapter #2
P3
P4
L1
L2
L3
L4
L5
L6
L7
L8
L9
L10
Switch S1
Switch S2
Switch S3
P6
Channel Adapter #4
P7
I/O controller
I/O controller
616
618
620
I/O Enclosure
608
606
P5
Channel Adapter #3
614
Third Host

FIG. 6

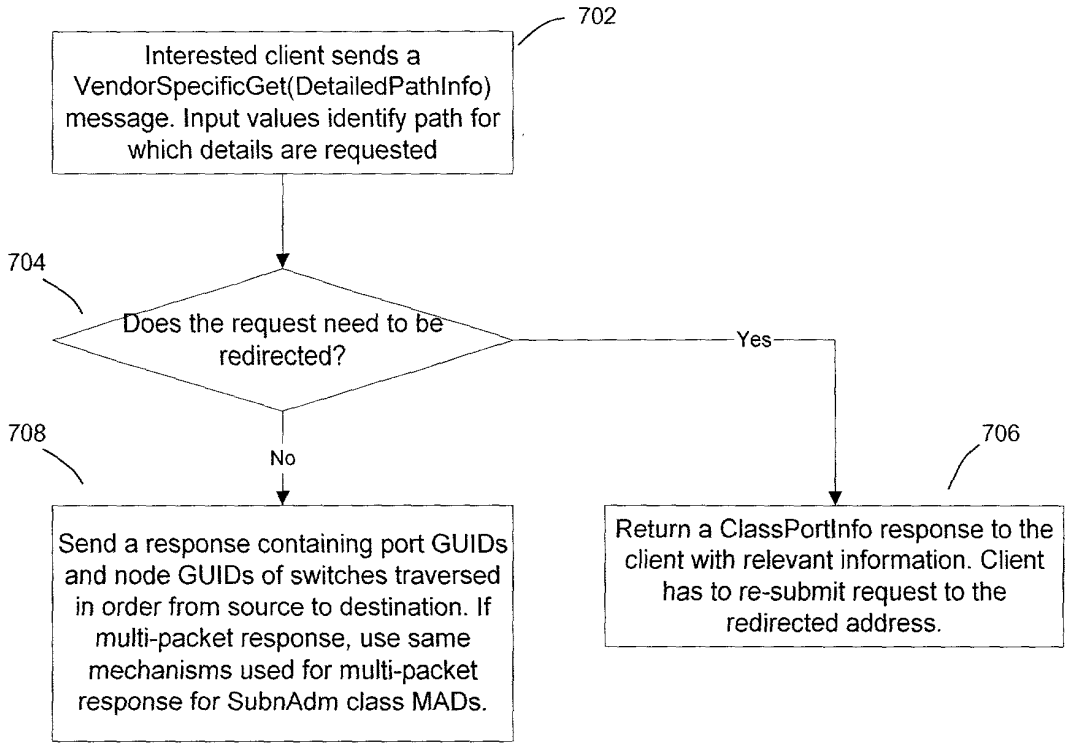

702
Interested client sends a VendorSpecificGet(DetailedPathInfo) message. Input values identify path for which details are requested
704
Does the request need to be redirected?
Yes
No
708
706
Send a response containing port GUIDs and node GUIDs of switches traversed in order from source to destination. If multi-packet response, use same mechanisms used for multi-packet response for SubnAdm class MADs.
Return a ClassPortInfo response to the client with relevant information. Client has to re-submit request to the redirected address.

FIG. 7

Attorney's Docket No.: 219.39026X00 (ATSK) PATENT
Intel No. P10093

DECLARATION AND POWER OF ATTORNEY FOR PATENT APPLICATION
(FOR **INTEL CORPORATION** PATENT APPLICATIONS)

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below, next to my name.

I believe I am the original, first, and sole inventor (if only one name is listed below) or an original, first, and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled SYSTEM AND METHOD FOR PROVIDING DETAILED PATH INFORMATION TO CLIENTS

the specification of which

XX is attached hereto.

___ was filed on _____ as
United States Application Number _____
or PCT International Application Number _____
and was amended on _____.
(if applicable)

I hereby state that I have reviewed and understand the contents of the above-identified specification, including the claim(s), as amended by any amendment referred to above. I do not know and do not believe that the claimed invention was ever known or used in the United States of America before my invention thereof, or patented or described in any printed publication in any country before my invention thereof or more than one year prior to this application, that the same was not in public use or on sale in the United States of America more than one year prior to this application, and that the invention has not been patented or made the subject of an inventor's certificate issued before the date of this application in any country foreign to the United States of America on an application filed by me or my legal representatives or assigns more than twelve months (for a utility patent application) or six months (for a design patent application) prior to this application.

I acknowledge the duty to disclose all information known to me to be material to patentability as defined in Title 37, Code of Federal Regulations, Section 1.56.

I hereby claim foreign priority benefits under Title 35, United States Code, Section 119(a)-(d), of any foreign application(s) for patent or inventor's certificate listed below and have also identified below any foreign application for patent or inventor's certificate having a filing date before that of the application on which priority is claimed:

Prior Foreign Application(s)			Priority Claimed	
(Number)	(Country)	(Day/Month/Year Filed)	Yes	No
(Number)	(Country)	(Day/Month/Year Filed)	Yes	No
(Number)	(Country)	(Day/Month/Year Filed)	Yes	No

I hereby claim the benefit under title 35, United States Code, Section 119(e) of any United States provisional application(s) listed below

(Application Number)	Filing Date
(Application Number)	Filing Date

I hereby claim the benefit under Title 35, United States Code, Section 120 of any United States application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States application in the manner provided by the first paragraph of Title 35, United States Code, Section 112, I acknowledge the duty to disclose all information known to me to be material to patentability as defined in Title 37, Code of Federal Regulations, Section 1.56 which became available between the filing date of the prior application and the national or PCT international filing date of this application:

(Application Number)	Filing Date	(Status -- patented, pending, abandoned)
(Application Number)	Filing Date	(Status -- patented, pending, abandoned)

I hereby appoint: Donald R. Antonelli, Reg. No. 20,296; David T. Terry, Reg. No. 20,178; Melvin Kraus, Reg. No. 22,466; William I. Solomon, Reg. No. 28,565; Gregory E. Montone, Reg. No. 28,141; Ronald J. Shore, Reg. No. 28,577; Donald E. Stout, Reg. No. 26,422; Alan E. Schiavelli, Reg. No. 32,087; James N. Dresser, Reg. No. 22,973; Carl I. Brundidge, Reg. No. 29,621; Paul J. Skwierawski, Reg. No. 32,173, my attorneys; of ANTONELLI, TERRY, STOUT & KRAUS, LLP with offices located at 1300 North Seventeenth Street, Suite 1800, Arlington, Virginia 22209, telephone: (703) 312-6600, fax: (703) 312-6666; and Alan K. Aldous, Reg. No. 31,905; Robert D. Anderson, Reg. No. 33,826; Joseph R. Bond, Reg. No. 36,458; R. Edward Brake, Reg. No. 37,784; Richard C. Calderwood, Reg. No. 35,468; Jeffrey S. Draeger, Reg. No. 41,000; Cynthia Thomas Faatz, Reg No. 39,973; Sean Fitzgerald, Reg. No. 32,027; Seth Z. Kalson, Reg. No. 40,670; David J. Kaplan, Reg. No. 41,105; Leo V. Novakoski, Reg. No. 37,198; Naomi Obinata, Reg. No. 39,320; Thomas C. Reynolds, Reg. No. 32,488; Steven P. Skabrat, Reg. No. 36,279; Howard A. Skaist, Reg. No. 36,008; Steven C. Stewart, Reg. No. 33,555; Raymond J. Werner, Reg. No. 34,752; and Charles K. Young, Reg. No. 39,435; my patent attorneys, and Calvin E. Wells, Reg. No. P43,256; and Alexander Ulysses Witkowski, Reg. No. P43,280; my patent agents, of INTEL CORPORATION; with full power of substitution and revocation, to prosecute this application and to transact all business in the Patent and Trademark Office connected herewith.

Send all correspondence to:

ANTONELLI, TERRY, STOUT & KRAUS, LLP
1300 North Seventeenth Street
Suite 1800
Arlington, VA. 22209

Direct all telephone calls and faxes to:

TEL: (703) 312-6600
FAX: (703) 312-6666

INTEL CORPORATION

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Full Name of Sole/First Inventor Rajesh Shah

Inventor's Signature Rajesh. Shah. Date Oct. 19, 2000

Residence 14320 N.W. Lilium Drive, Portland OR 97229 Citizenship Indian
(City, State) (Country)

Post Office Address Same as above

Full Name of Second/Joint Inventor _____

Inventor's Signature _____ Date _____

Residence _____ Citizenship _____
(City, State) (Country)

Post Office Address _____

Full Name of Third/Joint Inventor _____

Inventor's Signature _____ Date _____

Residence _____ Citizenship _____
(City, State) (Country)

Post Office Address _____

Full Name of Fourth/Joint Inventor _____

Inventor's Signature _____ Date _____

Residence _____ Citizenship _____
(City, State) (Country)

Post Office Address _____